Epiwork D3.1: Meta-model Initial Specification, Catalogue of Relevant Data, Platform Requirements

Luis F. Lopes[1], Fabricio Silva[1], Francisco Couto[1], Mario Silva[1]

[1] University of Lisbon, Faculty of Sciences, LASIGE, Portugal

30 September 2009 (revised 10 February 2010)

Abstract

This report introduces an information model for the epidemic marketplace, describes and discusses the architectural design of a metadata catalogue for the different kinds of datasets that may be included or referenced in the Epidemic Marketplace repository. Finally, it introduces the functional and non-functional requirements for the computational platform that will support the epidemic marketplace, including policies for uploading datasets and data harvesting.

Keyword List

Epidemic Marketplace, metadata catalogue, repository, dataset

Contents

List of Figures

1 Introduction

Epiwork proposes a multidisciplinary research effort, aimed at developing the appropriate framework of tools and knowledge needed for the design of epidemic forecast infrastructures, to be used by epidemiologists and public health scientists. The project is a truly interdisciplinary effort, anchored to the research questions and needs of epidemiology research by the participation in the consortium of epidemiologists, public health specialists, mathematical biologists and computer scientists.

The **Epidemic Marketplace (EM)** is the data management platform of Epiwork. The main components of the Epidemic Marketplace are:

1. A repository with epidemic data sets and a catalogue of epidemic data sources containing the metadata describing existing databases;
2. A forum to publish information about data, fostering collaboration among modellers;
3. Mediating software that can automatically process queries for epidemiological data available from the information sources connected to the platform.

The objectives of the Epiwork project where the Epidemic Marketplace will have a direct impact are:

1. Development of large scale, data driven computational models endowed with a high level of realism and aimed at epidemic scenario forecast.
2. Design and implementation of original data-collection schemes motivated by identified modelling needs, such as the collection of real-time disease incidence.
3. Setup of a computational platform for epidemic research and data sharing.

The objectives of this deliverable are:

- Introduction of the information model for the Epidemic Marketplace, including the main concepts to be used in its development.
- Presentation and discussion of the vision of an architectural design of a metadata catalogue for the Epidemic Marketplace
- Characterisation the information elements of the metadata catalogue in the context of the Epiwork project.
- Identification of the requirements of the Epidemic Marketplace to be implemented in Epiwork.

The catalogue, to be based on Semantic Web technologies, needs to be general enough to support the description of the different kinds of datasets that may be referenced or included in the Epidemic Marketplace repository. In addition, it should accept different levels of detail in metadata and be engineered to support its evolution, from a simple prototype only supporting free-text annotations in the initial stages, to a system where web resources can be fully described for automatic discovery and contents of datasets may be directly accessed.

1.1 Organization of the Report

This report is organized as follows:

Chapter 2, Information Model for the EM, introduces the information model, based on model-engineering concepts such as the use of metadata for epidemic resources characterization, that will be adopted in the Epidemic Marketplace.

Chapter 3, Ontologies, surveys the use of ontologies in the Biomedical domain, and their envisioned role in the Epidemic Marketplace.

Chapter 4, Metadata in the Epidemic Marketplace, discusses the strategy for modelling epidemic resources and their interlinking with increasingly higher levels of detail using existing W3C recommendations and other standards for using metadata.

Chapter 5, Catalogue, describes the requirements for modelling the metadata elements of the Epidemic Marketplace. These are derived from the analysis of a sample of epidemic datasets used in representative epidemic modelling studies.

Chapter 6, Platform Requirements, describes the general architecture of the EM and the various requirements, at the system and functional level, that have been collected,

Chapter 7, Conclusions and Future Work, outlines our plan for bringing up the Epidemic Platform within the Epiwork consortium and the community at large.

2 Information Model for the EM

To manage the information in the Epidemic Marketplace, mainly catalogues of datasets and the datasets themselves, it is necessary to adopt a common reference model and provide its description as metadata. Metadata is information about data. It provides a context for the data, helping to understand, to manage and to search it. The level of detail of metadata can change according to end use of the described data.

Metadata enables more correct and accurate data exchange and retrieval. The use of metadata standards makes the datasets' information models easier to be understood and used by different users and applications. As automatic tools for the manipulation, edition and exchange of data become more common and data needs to be machine-readable, the implementation of standard metadata becomes more and more important.

For example in the epidemic marketplace, the existence of metadata and a catalogue allows for the search of specific information without having to download and open a document to see its contents. If a researcher is looking for datasets relative to a specific disease or a specific geographic location, it is possible to obtain that information by searching the catalogue of metadata in the repository.

To build a data model, it is first necessary to identify the model, being that the basic things whose information needs to be stored and managed (Hay 2006). For example, for an epidemiological dataset, which can contain information about the number of people that have been infected with a specific disease, the instance data itself will define the model. These data need to be described with metadata. When we describe an Entity Class, such as *Patient*, which contains specific information about infected individuals, that information is metadata. We can also define in metadata the attributes that the *Patient* entity class can have, such as "name", "age", "gender" and so on.

Metadata can be specified at higher levels of abstraction. We can have meta-metadata, which describes and defines the metadata elements. In the example above, we the metadata model is composed by elements that are the *Entity Class* or its *Attributes*.

The use of metadata for the description of health related documents is, as in other areas, essential for the management of information and to keep data consistent. The use

of metadata facilitates communication and interoperability in electronic health data exchange, allowing a better standardization and data sharing among different services and even between different countries.

2.1 Ontologies in Metadata

One of the driving forces for the implementation of metadata and metadata standards is the W3C initiative for the Semantic Web (Feigenbaum et al. 2007). The use of metadata is fundamental for the development of the Semantic Web, where metadata annotated with ontologies will be essential for the development of machine-readable information.

Ontologies are formal systems of concepts intended to rigorously define what things mean and how they relate. An ontology is *an explicit specification of a conceptualization* (Gruber 1993).

The use of metadata to describe data and ontologies to describe relationships between data is becoming common practice in information and knowledge management. Metadata and ontologies are tools that can be applied to documents and other data sources that allow querying the underlying data sources in a more sophisticated, structured and meaningful manner. The use of controlled languages, such as ontologies, is essential for the description of data, maintaining metadata consistent.

For example, using a specific ontology to describe a specific disease makes everybody referring to a specific disease to use the same term, making the information discovery simpler and more complete. But it also keeps the metadata text simpler, since the ontology itself contains other data that doesn't need to be inserted as metadata. For example, through an ontology of places (a geographic ontology), if we have a specific location code, we can obtain other information about that location, such as country, coordinates, altitude, city and so on.

Chapter 3 surveys the use of ontologies in biomedical settings.

2.2 Metadata standards

There are several standards for the collection and management of metadata. ISO/IEC 11179 is the international standard for representing metadata for an organization in a Metadata Registry that has been implemented by organizations in the Health domain.

Perhaps a more relevant standard is Dublin Core (DC), which was conceived for describing web resources. The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description on the web. DC is of particular relevance to Epiwork, given that the information and computational platforms to be developed are intended to strongly adhere to Web standards.

We will survey both in the remainder of this section.

2.2.1 ISO/IEC 11179 Metadata Registry (MDR) Standard

The ISO/IEC 11179 is a standard for storing organizational metadata in a controlled environment, called a metadata registry. An ISO metadata registry consists of a hierarchy of *concepts* with associated properties for each concept. Here, concepts are similar to classes in object-oriented programming, but without the behavioural elements. Properties are similar to Class attributes.

The ISO/IEC 11179 MDR was designed for the use in enterprise. Several healh organizations are known to implement this MDR, such as the Australian Institute of Health and Welfare - Metadata Online Registry (METeOR, 2009); the US Health Information Knowledgebase (USHIK, 2009) and the US National Cancer Institute - Cancer Data Standards Repository (caDSR) (NCI Wiki, 2009).

The caDSr, developed by the US National Cancer Institute (http://www.cancer.gov/) has the goal of defining a comprehensive set of standardized metadata descriptors for cancer research data, both for information collection and analysis.

2.2.2 Dublin Core

The Dublin Core Metadata Element Set is a vocabulary of fifteen properties to be used to describe document-like files in the web. Those fifteen elements are a part of a

larger set of metadata vocabularies, the DCMI Metadata Terms [DCMI-TERMS], and technical specifications maintained by the Dublin Core Metadata Initiative (DCMI) (DCMI usage board, 2008).

Changes to the DC Metadata Element Set are regulated by the DCMI Namespace Policy [DCMI-NAMESPACE], which describes how DCMI terms are assigned Uniform Resource Identifiers (URIs) and sets limits on the range of editorial changes that may allowably be made to the labels, definitions, and usage comments associated with existing DCMI terms.

The fifteen element descriptions that have been formally endorsed in the ISO Standard 15836-2003 of February 2003, ANSI/NISO Standard Z39.85-2007 of May 2007 and IETF RFC 5013 of August 2007.

Since January 2008, DCMI includes formal domains and ranges in the definitions of its properties. This means that each property may be related to one or more classes by a *has domain* relationship, indicating the class of resources that the property should be used to describe, and to one or more classes by a *has range* relationship, indicating the class of resources that should be used as values for that property (Powell et al. 2008).

In order to not affect the conformance of existing implementations in RDF, domains and ranges have not been specified for the fifteen properties of the dc: namespace (http://purl.org/dc/elements/1.1/). Rather, fifteen new properties with "names" identical to those of the Dublin Core Metadata Element Set Version 1.1 have been created in the dcterms: namespace (http://purl.org/dc/terms/). These fifteen new properties have been defined as sub-properties of the corresponding properties of DCMES Version 1.1. The use of the new and semantically more precise dcterms is recommended in order to best implement the use of machine processable metadata.

2.2.3 Ontologies in Dublin Core

The DCMI recommends the use of controlled languages whenever possible for the description of each element. The use of specific ontologies can make the annotation more exact. Besides ontologies, other controlled languages can be used such as thesauri, A Thesaurus is not as complete as an ontology, but can be extremely useful for data standardization.

For example, there isn't a world geographic ontology available. However, the DCMI suggests the use of the TGN - Thesaurus of Geographic Names (Harpring 1997). Another example is specification of the type of resource, where it is recommended the use of the MIME codes (IAMA 2009).

The development of an ontology that is accepted by the whole community is a complex endeavour. One example of a successful case is the GO - Gene Ontology, which was developed with the involvement of the community.

In the Epidemic Marketplace, we intend to adopt controlled languages, such as the ones recommended by de DCMI (Dublin Core Metadata initiative) and others, such as the UMLS metathesaurus, which is a popular ontology for the biomedical domain, in metadata descriptions based on the Dublin Core Standards.

2.3 Open Archives Initiative

The Open Archives Initiative (OAI) develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content supported by open access movement (http://www.openarchives.org/).

The OAI has two ongoing projects, the OAI-PMH (Protocol for Metadata Harvesting) and the OAI-ORE (Object Reuse and Exchange).

OAI-PMH (2008) is useful for Dublin-Core metadata exchange using web protocols. An Epidemic Marketplace mediator service for accessing its metadata is likely to implement this standard.

OAI-ORE (2008) defines standards for describing and exchanging aggregated web resources. These aggregations may combine resources of different types into compound digital objects. This characteristic may be useful in the Epidemic Marketplace for creating compound objects from related datasets, better describing their relations and organizing them according to those relations.

3 Ontologies

Epidemiological research generates a vast amount of information that is ultimately stored in scientific publications or in databases. The information in scientific texts is unstructured and thus hard to access, whereas the information in databases, although more accessible, often lacks in contextualization. The integration of information from these two kinds of sources is crucial for managing and extracting knowledge. By structuring and defining the concepts and relationships within a domain, ontologies have taken a key role in this integration.

The use metadata to describe data and ontologies to describe relationships between data is being increasingly used in information and knowledge management. Metadata and ontologies can be applied to documents and other data sources that allow querying the underlying data sources in a more sophisticated, structured and meaningful manner. The use of ontologies becomes essential for the description of data, maintaining metadata consistent.

For example, the adoption of a specific ontology to describe a specific disease causes everybody referring to a specific disease to use the same term, making information discovery simpler and more accurate. It also keeps the metadata descriptions simpler, since the ontology itself contains other data that doesn't need to be inserted as metadata. For example, when using a geographic ontology, if we insert a specific geographic reference, such as a postal code, we can obtain from the ontology many other associated data, such as country, coordinates, altitude, and city, instead of inserting it directly as metadata.

This chapter describes the role of ontologies in sharing, integrating and mining epidemiological information, discusses some of the most relevant ontologies to Epiwork and illustrates how they are can used by the Epidemic Marketplace.

3.1 What is an Ontology?

Since Ancient Greece, philosophy has dealt with the need to define and structure reality. Aristotle proposed a system to organize the objects of human perception in well-

defined Categories, beginning with an explanation of synonyms, homonyms and paronyms. He recognized the importance of having clear unequivocal concepts to identify each object. In the 18th century, Linnaeus applied these same concepts to the natural world and developed a taxonomy for classification of living things. These early ideas have evolved into the current definition of Ontology in philosophy as a systematic account of Existence, and as such much more complex than Classification. Although the concept of Ontology has been in use by philosophy for a long time, it was only with the emergence of artificial intelligence that computer science borrowed the term to establish content-specific agreements for the sharing and reuse of knowledge among software systems. In this context, Gruber (1991) defines an ontology as a specification of conceptualisations, used to help programs and humans share knowledge. Conceptualisations refer to the entities: the terms, the relationships between them, and also the constraints of those relationships. On the other hand, specification refers to the explicit representation of the conceptualisations.

Using this general description, controlled vocabularies, taxonomies and thesaurus can be considered ontologies (Bodenreider & Stevens 2006). A **controlled vocabulary** is a list of terms that have been explicitly enumerated. A **taxonomy** is a collection of controlled vocabulary terms organised into a hierarchical structure. A **thesaurus** is a networked collection of controlled vocabulary terms.

Ideally, an ontology should contain formal explicit descriptions of the concepts (often called classes) in a given domain, which should be organized and structured according to the relationships between them. They also make the relationship between concepts explicit, which allows further reasoning and enables a fuller representation of the information by including such aspects as interacting partners, specific roles, and functions in specific contexts or locations.

According to Stevens et al. (2000), ontologies have been classified into three types:

1. **Domain-oriented**: either domain specific (e.g. ontology dedicated to a single disease) or domain generalisations (e.g. dedicated to European diseases);
2. **Task-oriented**: e.g. for clinical analysis;

3. **Generic**: defining high-level categories that are maintained across several domains (also called top-level or upper-level ontologies).

A well-structured ontology will reuse ontologies of the three types, but in a clearly defined modular way, to allow structural modifications and concept reusability.

The role of ontologies has changed in recent years: from limited in scope and scarcely used by the community, to a main focus of interest and investment. Although clinical terminologies have been in use for several decades, different terminologies were used for several purposes, hampering the sharing of knowledge and its reliability. This has lead to the creation of ontologies to answer the need to merge and organize the knowledge, and overcome the semantic heterogeneities observed in this domain. While the first attempts at developing them focused on a global schema for resource integration, real success and acceptance was only achieved later by ontologies for annotating entities (Bodenreider & Stevens 2006). Since then, ontologies have been used successfully for other goals, such as the description of experimental protocols and medical procedures.

The examples that follow provide an illustration of some of the most widely-used ontologies that could be adopted by the Epidemic Markteplace.

3.2 BioMedical Ontologies

The Unified Medical Language System (www.nlm.nih.gov/research/umls/) (UMLS) is a compendium (or an integrated ontology) of text mining-oriented biomedical terminology encompassing all aspects of medicine (Bodenreider 2004). It comprises three distinct knowledge sources: the *Metathesaurus*, the *Semantic Network*, and the *SPECIALIST lexicon*.

The *Metathesaurus* is an extensive, multi-purpose vocabulary database that integrates information from over one hundred clinical and biomedical databases and information systems, such as ICD, MeSH, SNOMED and GO. It defines biomedical concepts, listing their various names and relationships and mapping synonyms from different sources, thus providing a common knowledge basis for information exchange. It can be used autonomously for a variety of applications, namely linking between

different clinical or biomedical information systems, and linking patient records to literature sources and factual databases. However, its utility is enhanced when used with the other UMLS knowledge sources. Since the Metathesaurus contains concepts and terms from diverse sources for diverse purposes, many specific applications require a customized reduced version of it, where only the areas of interest are included.

The *Semantic Network* is an ontology of biomedical subject categories (*semantic types*) and relationships between them (*semantic relations*) with the purpose of semantically categorizing the concepts from the *Metathesaurus* (each term in the *Metathesaurus* is linked to at least one *semantic type*). *Semantic types* are organized in a tree-structure with major types including *organism*, *anatomical structure*, *biologic function*, *chemical*, and *event*. The tree edges are labelled with the main *semantic relation*, *is-a*, although several other non-hierarchical semantic relations also exist, grouped in five major categories: *physically related to*, *spatially related to*, *temporally related to*, *functionally related to*, and *conceptually related to*.

The *SPECIALIST Lexicon* is an English language lexicon focused on biomedical vocabulary, but also including common English words. Each entry in the lexicon, or lexical item, includes syntactic, morphological and orthographic information, essential for natural language processing (NLP). This lexicon was developed to support an NLP system, also called *SPECIALIST*, which is available with the UMLS as a set of *lexical tools*.

The UMLS was developed and is maintained by the US National Library of Medicine, with its main goal being the improvement of accessibility to biomedical information by facilitating its interpretation by computer systems. It successfully addresses the problem of coping with the multiplicity of vocabularies and terminologies in use in medicine through an integrative approach (the *Metathesaurus*) and complements it with a semantic structure that facilitates computer reasoning (the *Semantic Network*) and lexical information. This enables NLP-based text mining tools to explore the biomedical literature (the *SPECIALIST Lexicon*). These three factors, together with the all-encompassing scope of UMLS, make it an invaluable tool for mining medical data in any of its aspects.

Other ontologies exist in this area, for example SNOMED CT - Systematized Nomenclature of Medicine-Clinical Terms (Ahmadian et al. 2009) (http://www.nlm.nih.gov/snomed). SNOMED CT is a comprehensive clinical terminology that was formed by the merger, expansion, and restructuring of SNOMED RT (Reference Terminology) and the United Kingdom National Health Service (NHS) Clinical Terms (also known as the Read Codes). SNOMED CT is oriented to concepts using a well-formed, machine-readable terminology.

3.3 Spatial/geographic Ontologies

A geographic ontology describes spatial entities corresponding to features such as region boundaries or natural resource classifications.

A popular database of geographical names is GeoNames (Wick & Becker 2007) (http://www.geonames.org), which in September 2009 contained over 8 million geographical names and consisted of 6.5 million unique features. For instance, the geographical name *Sintra* can correspond to different features, such as populated place or a mountain. By using GeoNames we can disambiguate geographical names and obtain their exact location (latitude/longitude). The information can be freely obtained from the intranet using Semantic Web technology. The GeoNames Ontology (www.geonames.org/ontology) adds geospatial semantic information by describing and interlinking features. For each geospatial location the ontology provides its children, neighbors, or nearby locations.

Another example, is the full geographic ontology of Portugal GEONET-PT (http://xldb.fc.ul.pt/wiki/Geo-Net-PT_02), which was developed by the LASIGE group of the Epiwork Consortium. Geo-Net-PT contains more than 415 thousand features. Geo-NET-PT, which is released following the W3C recommendations for ontology representation (McGuinness et al. 2004), was generated with GKB, a common knowledge base for integrating data from multiple external resources (i.e. public gazetters and databases). GKB essentially manages place names and the ontological relationships between them (i.e. broader/narrower geographical entities), supporting mechanisms for storing, maintaining and exporting this information (Chaves et al. 2005).

An example of a commercial product is TGN, the Getty Thesaurus of Geographic Names (Harpring 1997), The J. Paul Getty Trust developed this product to provide terminology and other information about the objects, artists, concepts, and places important to various disciplines that specialize in art, architecture and material culture. TGN includes names and associated information about places, which may be administrative political entities (e.g., cities, nations) and physical features (e.g., mountains, rivers). Given its purpose TGN includes also relevant information related to history, population, culture, art and architecture.

As the previous projects show this is an emergent and important field. This motivated the European Commission to recently support the development of the Infrastructure for Spatial Information in Europe (INSPIRE) initiative (European Commission 2002). Besides promoting and facilitating the interchange of environmental spatial information among organizations, this initiative aims at providing easy and public access to spatial information across Europe through a single infrastructure.

3.4 The Role of Ontologies in the Epidemic Marketplace

The Epidemic Marketplace aims at providing a unified and integrated approach for the management of epidemic resources. To achieve this goal the Marketplace has to implement mechanisms to allow users to describe their data in a standard way requiring minimal human intervention. This can effectively be accomplished by the integration of established and comprehensive ontologies, not only in the acquisition of data from the users but also in the automatic retrieval and query of external data.

The use of biomedical ontologies, such as UMLS, to describe and understand epidemic datasets is obvious given their biomedical nature, but geospatial referenced data is also essential to epidemiologic studies. A geographic ontology, such as GeoNames or TGN, providing a coherent geospatial annotation of epidemic datasets will be crucial for an effective understanding of disease propagation and prediction. Therefore, we intend to keep track of the novel services provided by the European Commission initiative, INSPIRE, to integrate them in the Epidemic Marketplace.

At the first stage, the Epidemic Marketplace aims at creating a catalogue of epidemic datasets with extensive meta-data describing their main characteristics. Ontologies will play an important role in establishing the common terminology to be used in this process and to interlink heterogeneous meta-data classifications. The Epidemic Marketplace will explore not only one ontology but a comprehensive set of relevant ontologies that besides being used to characterise datasets will also become important datasets to epidemic modellers. Some of these are already being organised in collections. OBO (The Open Biomedical Ontologies) is a repository of many relevant ontologies to Epiwork, openly available from http://www.obofoundry.org/.

At a later stage, the marketplace will provide a unified and integrated approach for the management of epidemic data sources. Ontologies will have an important role in integrating these heterogeneous data sources by providing semantic relationships among the described objects. Further on, the marketplace will include methods and services for aligning the ontologies. The aligned ontologies and annotated datasets will eventually serve as the basis for a distributed information reference for epidemic modellers, which will help further on the integration and communication among the community of epidemiologists.

4 Metadata in the Epidemic Marketplace

A key component of the Epidemic Marketplace platform is a semantically enabled repository. The prime objective of the EM repository is to organize epidemic or related information in the form of datasets and/or their metadata. Epidemiological datasets may contain different types of data, which may be useful for the understanding of epidemics and disease propagation, from disease data to geographic or demographic data that can be used for modelling disease transmission or statistical analysis.

The objective of the Epidemic Marketplace repository is to organize the information about existing datasets. While it is expected that the datasets are deposited in the repository, it is possible to have information about specific datasets even if they are not stored at the repository. This may happen, for example, for security reasons.,

For these special datasets, the metadata services to be provided by the content repository will become the only alternative. The metadata repository will store information about specific datasets even if they are not in the repository. The metadata will describe the datasets in detail, including their contents, providing information about the authors, where the dataset is available and who has access to it.

To organise and manage the metadata, a catalogue will be produced, that will enable a faster and more accurate search of specific epidemiological information. The management of resources by the repository and its metadata will be done at different levels:

1. **Resource level**, where the dataset will be described using properties with the semantics of the DC elements defined for that purpose;
2. D**omain level,** to describe the contents of the datasets. This will be done by the use of properties with the semantics of DC elements defined for that purpose, encoded with extensions to be proposed by the Epiwork repository application profile being elaborated;
3. **User level**, where users who access a resource will have the possibility of commenting and leaving information about that dataset, possibly when deriving new datasets from that resource, which will then be also annotated with metadata.

4.1 Epidemic Resources Metadata

To describe the epidemic datasets, it is first necessary to describe the dataset as a web resource. This will be done using the DCMI terms and conventions, but it also necessary to describe the information contained in the datasets. These descriptions constitute what health professionals and researchers will be ultimately looking for.

To describe the contents of epidemic and other related datasets it is necessary to propose a general metadata model capable of describing virtually every kind of information, given the diversity of factors and the interdisciplinary of epidemiologic studies. In the study of a specific disease it is possible to have datasets describing the disease, how it spreads, clinical data about a population and so on. Data may be geo-referenced and geographic data may be necessary for the modelling of the disease transmission. Other data can be important for the study of diseases, such as genetic, socio-economic, demographic, environmental and behavioural data. The need to encompass so many areas of study will reflect on the contents of the datasets and ultimately on their metadata.

The level of detail of the metadata is however something that must be carefully designed: a low level of detail may not be able to sufficiently describe the datasets, making the right information harder to find, but a too detailed metadata scheme can turn the annotation of a specific dataset into a daunting task, hindering the acceptance of the model by the user community.

In view of this we intend to start modelling the datasets with a low level of detail, annotating the 15 standard DC elements as character data (see Figure 1); further down the line, we will support the extension of the DC elements annotations with semantically richer descriptions. That will be initially done with the analysis of datasets to be provided by Epiwork partners. The collaboration with these partners will enable the assessment of which level of detail will be most adequate to the epidemic modellers community.

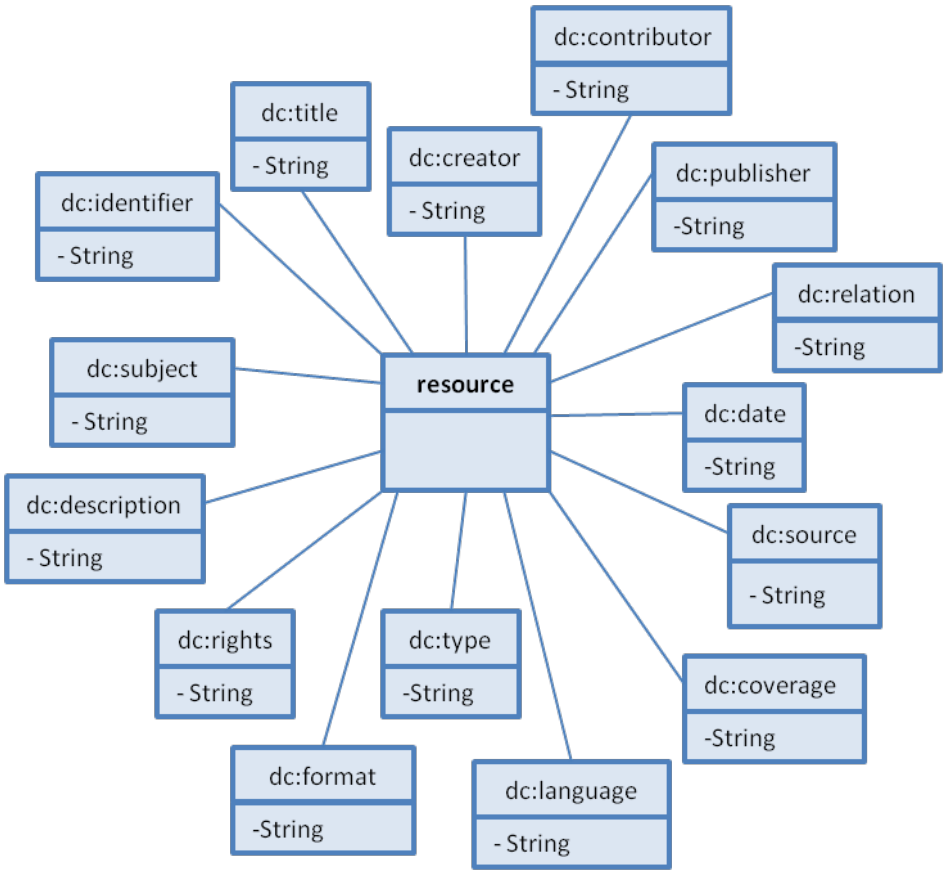

Figure 1 - The 15 Dublin Core Elements.

For the metadata annotation to be useful it needs to be annotated in a standard way, so data can be comparable and searched using similar queries. In order to obtain a standardization of the metadata annotation it is fundamental to use controlled languages as much as possible and languages for describing data structures, progressively limiting the use of free text. The annotation of metadata with free text is not recommended since it is not amenable to automatic processing, and it is subjective, leading to different people annotating the same dataset using different terms and different levels of detail.

There are several proposals to manage this issue. For example, for representing dates, it is recommended the use as the format defined in a profile of ISO 8601, an international standard to represent dates and times (http://www.w3.org/TR/NOTE-datetime), such as YYYY-MM-DD format.

There are controlled languages that can be adopted and will become fundamental for the standardization of epidemiological models, such as the UMLS metathesaurus (Bodenreider 2004). For instance, we intend to use UMLS to code diseases, thus avoiding the use of different terms to refer to the same disease. The same is applied to other types of data, such as geographical data, for which a world geographic ontology or thesaurus such as the TGN (Thesaurus of geographic names) may be used (Harpring, 1997).

5 Catalogue

The search of information in a repository, especially if the repository has a large number of datasets, can be a hard task to complete successfully. The difficulty in easily retrieving the relevant datasets would be a major setback for the implementation and acceptance of such platform by the user community.

In order to overcome this issue the Epiwork information platform will include a metadata catalogue that will support accurate searches for epidemic datasets.

The implementation of this catalogue will be phased. At first a simple Dublin Core (DC) scheme with the 15 legacy DCMI elements will be used in order to annotate the datasets (Dublin Core 2009). Later, a metadata schema for epidemiologic and related datasets will be developed based on the current DCMI terms (Dublin Core 2009) and Epiwork extensions.

That metadata modelling will be based mostly on the analysis of different datasets that should be identified and provided by the Epiwork consortium partners.

In order to understand the metadata to be added to annotate epidemic datasets and what properties should be extended in the future for a better data representation, we have analysed a selected sample of datasets:

EM Twitter Datasets: Twitter data harvested by an initial prototype of the Data Collector module of the Epidemic Marketplace (Lopes et al. 2009)

US Airports Dataset: Data about the airport network of the United States (supplied by ISI)

In addition, to add more diversity to these initial datasets and start with a larger study base, we surveyed published articles in epidemiology journals for analysis and inferred the attributes of that datasets reported in those papers. Most of the studies do not provide information on how to access all the used datasets or fully describe them for the purposed of cataloguing with the detail we are envisioning for the EM. Nevertheless, this kind of survey provides insights on the metadata modelling aspects that the EM should support to address the requirements of Epiwork.

We characterized datasets used in three studies,

Cohen et al. (2008) – analyses the relation of levels of household malaria risk with topography related humidity.

East et al. (2008) - analyzes the patterns of bird migration in order to identify areas in Australia where the risk of avian influenza transmission from migrating birds is higher.

Starr et al. (2009) - introduces a model for predicting the spread of *Clostridium difficile* in hospital context.

Using this approach, we have annotated several datasets, to which we did not actually have access, but devised what would be their metadata description as DC elements, based on the information provided.

We now present the DC metadata for annotating the above datasets, given as examples of the kind of metadata that will be available in the EM Metadata Repository.

5.1 EM Twitter datasets

These datasets contain Twitter messages mentioning disease names and location keywords. The Twitter datasets were produced using an initial prototype of the Data Collector of the Epidemic Marketplace (Lopes et al. 2009).

Each dataset contains tweets (messages) with disease and geographic specific keywords. It also contains, for each message, information about the author name (nickname), the source (in this case the Twitter.com service), the keywords searched, the date, the source and a possible score (assigned according to the confidence on the specific message).

A simplified XML, with the 15 DC elements filled-in to appropriately characterize a dataset containing information relative to Twitter messages containing the words Portugal and H1N1 (and aliases) in the text body, is given in Figure 2 in order to introduce the kind of annotations that will be stored in the Catalogue using the simple DC schema. These DC elements capture information that needs to be known by the

users of the dataset. Their contents are given as free text, but, as discussed in Chapter 4, they should evolve to machine-readable dcterms in the future.

```
<dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"  xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance">
        <dc:contributor>Luís Filipe Lopes</dc:creator>
        <dc:contributor>Joao M Zamite</dc:contributor>
        <dc:contributor>Bruno C Tavares</dc:contributor>
        <dc:contributor>Francisco M Couto</dc:contributor>
        <dc:contributor>Fabricio Silva</dc:contributor>
        <dc:contributor>Mario J Silva</dc:contributor>
        <dc:coverage>Spatial:Portugal</dc:coverage>
        <dc:coverage>Temporal: 16-5-2009 to 3-6-2009</dc:coverage>
        <dc:language>English</dc:language>
        <dc:language>Portuguese</dc:language>
        <dc:source>http://epiwork.di.fc.ul.pt/collector/</dc:source>
        <dc:identifier>dataset-twitter-003</dc:identifier>
        <dc:format>text/tab-separated-values</dc:format>
        <dc:date>2009-05-29</dc:date>
        <dc:title>Twitter dataset H1N1 + Portugal 4-6-2009</dc:title>
        <dc:creator> LASIGE node of the Epidemic Marketplace</dc:creator>
        <dc:subject>twitter message dataset</dc:subject>
        <dc:type>dataset</dc:type>
        <dc:description> This dataset contains Twitter messages containing the words H1N1 and Portugal
        collected between 16-5-2009 and 3-6-2009,
        Information is a 7 columns relation, containing the following data:
        Column 1- keyword 1 (disease)- H1N1
        Column 2- Keyword 2 (location)- Portugal
        Column 3- Source (Twitter)
        Column 4- Author of the message (user id)
        Column 5- The message body (evidence)
        Column 6- score
        Column 7- date (day and hour)</dc:description>
        <dc: publisher >Epiwork – http://www.epiwork.eu </dc:publisher>
        <dc:relation> Luis F. Lopes, João M. Zamite, Bruno C. Tavares, Francisco M. Couto, Fabrício
        Silva and Mário J. Silva. (2009). Automated Social Network Epidemic Data Collector. INForum
        informatics symposium.</dc:relation>
        <dc:rights> Creative Commons Attribution-ShareAlike (CC BY-SA),
        http://creativecommons.org/licenses/by-sa/3.0/
        </dc:rights>
</dc:dc>
```

Figure 2- Proposed DC elements for an EM Twitter dataset.

The **dc:creator** element is indicates the author of the dataset. The creator may be a person or an Institution, being in this example LASIGE, one of the Institutions participating in the Epiwork project.

The **dc:publisher** should refer to the organization/institution that issues the resource, in this case it is set as Epiwork, the project in which context this dataset has been produced.

The **dc:coverage** describes the scope of the dataset. In this case, the dataset is relative to Portugal. With the adoption of a geographic controlled vocabulary, such as TGN or GeoNames, this field will be annotated according to specific codes of the vocabulary. The period when the observations took place is also very relevant in epidemic studies. As a result, the coverage should also describe the temporal scope, so this is a property that could be extended in order to describe both concepts. A specific format should be implemented in order to distinguish spatial from temporal coverage. In the example, the coverage is informally annotated as: Spatial:<location> and Temporal:<time>.

The **dc:description** element is given as free text. In the future, it is desirable to extend the metadata model, so information that is now in this field could be placed in specific fields, using ontologic terms. For instance, the schema of the dataset, if it is published as XML data, could be described for instance in XSD (http://www.w3.org/XML/Schema).

The **dc:language** element is used describe the languages in a resource. As this contains Twitter messages in English and Portuguese, both languages are present.

The **dc:source** represents where the dataset was originated. In this case the source is the Epidemic Market Place Data Collector, which obtained the information from Twitter.

The **dc:identifier** is reserved for the indication of string, such as an ISBN or a DOI, identifying the resource. For this dataset, we specified its URL in the initial prototype of the Epidemic Marketplace under development (http://epiwork.di.fc.ul.pt)

The **dc:format** should be described using a standard such as the MIME media type standard to describe file contents in the web. In this case the MIME type is "text/tab-separated-values" since the dataset is made available as a text file in the TSV (tab separated values) format.

The **dc:date** element should be used to annotate the date at which the dataset was created or last modified. It should be annotated according to a standard. In this case the ISO 8601 (ISO 8601:2004), which is an international standard for the exchange of date and date related data. The date should follow this standard being annotated in the YYY-MM-DD format.

The **dc:title** field is a text field where the title of the resource should be provided.

The **dc:subject** field should be used to describe what is the subject of the dataset. In this dataset, here we indicate what the type of information the dataset contains, temporal coverage and describe the columns of the dataset.

The **dc:type** field is used to indicate the type of the resource. In this case it is a dataset. This field should be extended in order to describe in more detail the dataset using ontologic terms, indicating what is the specific type of the dataset, for example if it is an epidemiological, clinicial or geographic dataset.

The **dc:relation** field is used to refer other information that may be related with this specific one. A dataset can be referred or an article such as in this example.

The dc:rights field should be used for the description of who owns a resource and what kind of access is given to other users. In this dataset, we have indicated the Creative Commons Attribution-ShareAlike (CC BY-SA) license (http://creativecommons.org/).

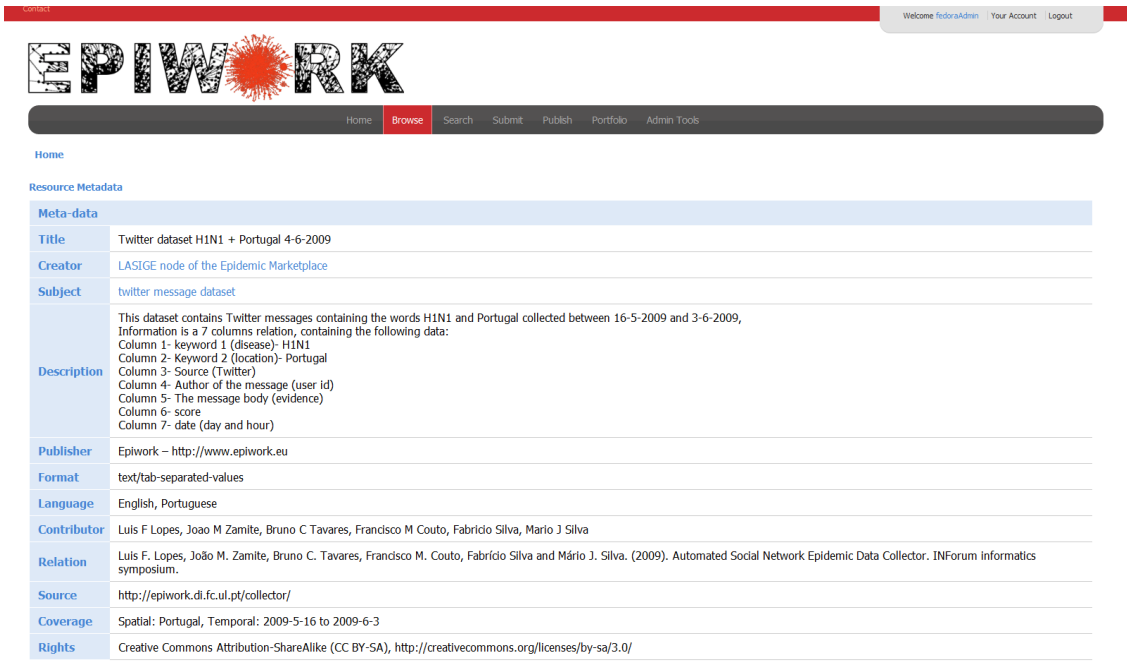

Figure 3- Gives an illustration of how the DC elements of this example dataset are displayed and edited in a first prototype of the catalogue now under development.

5.2 US Airports Dataset

This dataset provides information about the US transportation network, containing data about the 500 US airports with most traffic. The file contains an anonymized list of connected pairs of nodes and the weight associated to the edge, expressed in terms of number of available seats on the given connection on a yearly basis. Figure 4- Proposed DC elements for an EM US airport datasetrepresents the DC elements that we have associated to this dataset in the initial prototype of the EM under development.

Next, we discuss the new issues that are brought by the analysis of the DC elements proposed for this dataset, since each field was individually addressed in the previous dataset and most of the considerations are similar here. It is worth to note that some of the DC elements are have not been filled for this dataset:

```
<dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
        <dc:contributor> Daniela Paolotti, ISI </dc:creator>
        <dc:coverage>United States</dc:coverage>
        <dc:language> </dc:language>
        <dc:source> </dc:source>
        <dc:identifier></dc:identifier>
        <dc:format>text/plain</dc:format>
        <dc:date>2009-09-03</dc:date>
        <dc:title> US Air Transportation Network </dc:title>
        <dc:creator> ISI node of the Epidemic Marketplace</dc:creator>
        <dc:subject> Undirected weighted network of the 500 US airports with the largest amount of
        traffic </dc:subject>
        <dc:type>dataset</dc:type>
        <dc:description> </dc:description>
        <dc:publisher >Epiwork – http://www.epiwork.eu </dc:publisher>
        <dc:relation></dc:relation>
        <dc:rights> Please, feel free to use the above network dataset, provided the appropriate credit is
        given to the authors </dc:rights>
</dc:dc>
```

Figure 4- Proposed DC elements for an EM US airport dataset

The **dc:identifier** is not known, but the URL or a DOI handle pointing to the copy of this dataset in the prototype could be provided.

The **dc:relation** is empty, because we don't know where it is officially described

The **dc:source** is empty for the same reason.

The **dc:language** is empty (could be undefined), because language does not apply, as the dataset only has numbers, according to the EMPTY description

5.3 Cohen et al. (2008) dataset

The article by Cohen et al. (2008), presents a work where levels of household malaria risk are related with topography related humidity.

In our reading of this article, four distinct datasets have been indentified:

1. One that contains geographic data, such as topological maps;
2. One that contains socio-demographic data, such as mortality and birth rates;

3. An environmental dataset, containing humidity data for locations;
4. a clinical/epidemiological dataset, that should contain information about the pathogen, vector, diagnostics, etc.

The fourth dataset described in this article gives an example of a clinical/epidemiological dataset containing data about the disease, vector and infection diagnostics.

Next we discuss the new issues that are brought by the analysis of the DC elements proposed for this dataset, given in Figure 5, since each field was individually addressed in the previous dataset and most of the considerations are similar here.

The **dc:identifier** was left empty, because we don't know if and where the data is available. It is possible that this identified dataset is composed of data obtained from more than one more actual dataset. For example diagnostic results, with the specific parasite found, could be stored in one dataset, information about vectors species identified and parasites they're infected with in another one. In that case, the lineage of the data should also be recorded in the DC elements.

The **dc:creator** is defined as the first author of the article, because we are assuming it.

The **dc:contributor** is defined as "co-workers", since we don't really know exactly who worked on the production of the dataset and we're assuming it for this exercise.

The **dc:publisher** is empty because it is unknown

The **dc:rights** is empty as well.

The **dc:rights** is empty, because we don't know about the availability of the data.

```
<dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <dc:contributor>coworkers</dc:contributor>
    <dc:coverage>Western Kenya Highlands</dc:coverage>
    <dc:language>English</dc:language
    <dc:source> </dc:source>
    <dc:identifier></dc:identifier>
    <dc:format>text /plain</dc:format>
    <dc:date>2008</dc:date>
    <dc:title>supposed-dataset-cohen-et-al-2008-03</dc:title>
    <dc:creator>Cohen</dc:creator>
    <dc:subject> malaria epidemiology</dc:subject>
    <dc:type> dataset</dc:type>
    <dc:description> infected, vector, diagnostic </dc:description>
    <dc: publisher > </dc:publisher>
    <dc:relation>article: Cohen, J.M, Ernst, K.C., Lindblade K.A., Vulule J.M., John C.C. and
    Wilson M.L. (2008). Topography-derived wetness indices are associated with household-level
    malaria risk in two communities in the western Kenyan highlands. Malaria J., 7: 40.
    </dc:relation>
    <dc:rights></dc:rights>
</dc:dc>
```

Figure 5- Proposed DC elements for an Cohen et al. (2008) dataset.

The **dc:source** is empty, because we don't know where the dataset can be obtained.

The **dc:date** only specifies the year (we are assuming that this dataset was created in the year when the article describing it was published).

5.4 East et al. (2008) dataset

In this study the authors analyze patterns of bird migration in order to identify areas in Australia where the risk of avian influenza transmission from migrating birds is higher. Several datasets used in this study can be identified:

1. geographic datasets, with maps, aereal photos, addresses, etc;
2. epidemiological datasets, comprising data from the analysis of bird infections;
3. bird demographic datasets, with bird densities and distribution.

A possible annotation of the third dataset derived from this article is given in Figure 6. We are considering here a geographic dataset that could be formed by a set of map and aerial photography images in jpeg format, for example.

```
<dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <dc:contributor>coworkers</dc:contributor>
    <dc:coverage>Australia</dc:coverage>
    <dc:language>English</dc:language>
    <dc:source>.</dc:source>
    <dc:identifier>supposed-dataset-article-005</dc:identifier>
    <dc:format>image/jpeg</dc:format>
    <dc:date>2008 </dc:date>
    <dc:title>supposed-dataset-east-et-al-2008-01</dc:title>
    <dc:creator>East</dc:creator>
    <dc:subject> maps and aerial photos</dc:subject>
    <dc:type> dataset</dc:type>
    <dc:description> maps, aerial photos </dc:description>
    <dc: publisher > </dc:publisher>
    <dc:relation>article: East I.J., Hamilton S. and Garner M.G. (2008). Identifying areas of
    Australia at risk of H5N1 avian influenza infection from exposure to migratory birds: a spatial
    analysis. Geospatial health 2(2):203-213.</dc:relation>
    <dc:rights></dc:rights>
</dc:dc>
```

Figure 6- Proposed DC elements for East et al. (2008) dataset.

No new issues have been encountered in this dataset. All the discussions about how to fill the DC elements of this dataset were similar to those found in the annotation the datasets described earlier.

5.5 Starr et al. (2009) dataset

The article introduces a model for predicting the spread of *Clostridium difficile* in hospital context.

Several datasets used in this study can be indentified:

1. clinical datasets, with data from infected patients;
2. epidemiological datasets, containing data about transmission variables used to define the transmission model.

A possible annotation of the third dataset derived from this article is given in Figure 7. We are considering here an epidemiological dataset, containing information about transmission of the disease that was used for the development of the model presented in this article. This dataset could contain data know from other previous studies, for the calculation of specific parameters to be used in the model as well as data derived from the clinical datasets.

```
<dc:dc xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <dc:contributor>coworkers</dc:contributor>
    <dc:coverage> Hospital interior</dc:coverage>
    <dc:language>English</dc:language>
    <dc:source> Starr J.M., Campbell A., Renshaw E., Poxton I.R., and Gibson G.J. (2009).
    Spatio-temporal stochastic modelling of Clostridium difficile. The Journal of Hospital
    Infection 71(1):49-56.</dc:source>
    <dc:identifier>supposed-dataset-article-010</dc:identifier>
    <dc:format>text/plain</dc:format>
    <dc:date>2008</dc:date>
    <dc:title>supposed-dataset-starr-et-al-2009-01</dc:title>
    <dc:creator>Starr</dc:creator>
    <dc:subject>clostridium dificile clinical infection data</dc:subject>
    <dc:type> dataset</dc:type>
    <dc:description>transmission parameters used for the development of the transmission model
    </dc:description>
    <dc: publisher > </dc:publisher>
    <dc:relation>article: Starr J.M., Campbell A., Renshaw E., Poxton I.R., and Gibson G.J.
    (2009). Spatio-temporal stochastic modelling of Clostridium difficile. The Journal of Hospital
    Infection 71(1):49-56.</dc:relation>
    <dc:rights></dc:rights>
</dc:dc>
```

Figure 7- Proposed DC elements for Starr et al. (2009) dataset.

The problems encountered here were similar to the ones in the sections 5.3 and 5.4. Also a note about the spatial coverage, since the model developed is relative to transmission of *Clostridium difficile* in a general hospital seting, rather than a specific geographic location.

6 Platform Requirements

The Epidemic Marketplace can be defined as a *distributed virtual repository,* a platform supporting *transparent*, seamless access to distributed, heterogeneous and redundant resources (Kuliberda et al. 2006, Ohno-Machado et al. 1997)

It is a *virtual* repository because data can be stored in systems that are external to the Epidemic Marketplace, and it provides *transparent* access because several heterogeneities are hidden from its users.

This chapter describes the General Architecture that we propose for the Epidemic Marketplace and its various requirements:

- System Requirements
- Hardware Requirements for the System Implementation
- Non-functional Requirements
- Functional Requirements

6.1 Epidemic Marketplace General Architecture

The Epidemic Marketplace is composed of a set of interconnected data management nodes geographically distributed, sharing common canonical data models, authorization infrastructure and access interfaces. Data can be either stored in one or more repositories or retrieved from external data sources using authorization credentials provided by clients. Data can also be replicated among repositories to improve access time, availability and fault tolerance. However, data replication is not mandatory; in several cases data must be stored in a single site due to, for instance, security constraints. It is worth noting, though, that any individual repository that composes the Marketplace will enable virtualized access to these data, once a user provides adequate security credentials.

As shown in Figure 8 each Epidemic Marketplace node has the following modules:

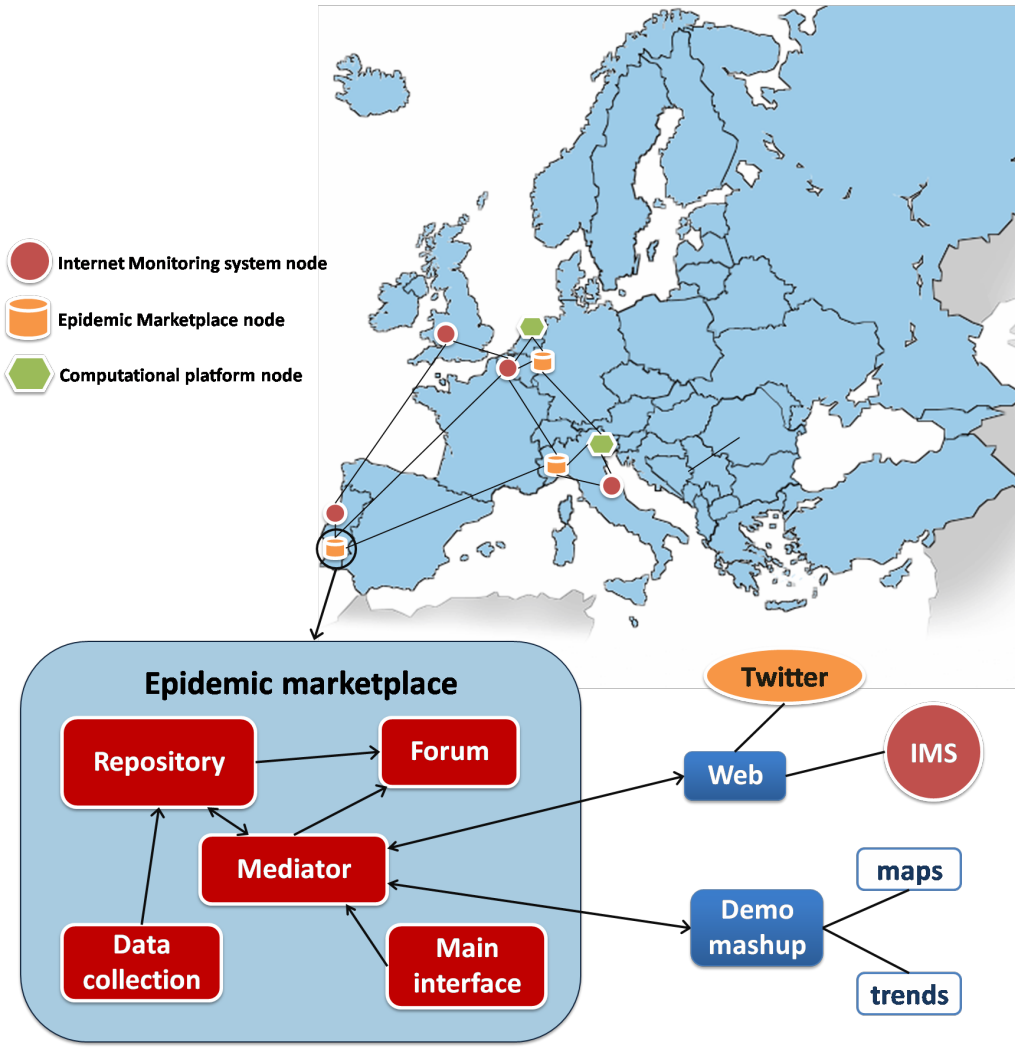

Figure 8 - An envisioned deployment of the distributed Epidemic Marketplace.

Repository: stores epidemic data sets and an epidemic ontology to characterise the semantic information of the data sets.

Mediator: a collection of web services that will provide access to internal data and external sources, based on a catalogue describing existing epidemic databases through their metadata using state-of-the-art semantic-web/grid technologies.

Collector: retrieves information of real-time disease incidences from publicly available data sources, such as social networks; after retrieval, the collector groups the incidences by subject and creates data sets to store in the repository.

Forum: allows users to post comments on integrated data from other modules, fostering collaboration among modellers;

6.2 System requirements

Several requirements have been identified when defining the architecture of the Epidemic Marketplace. Those requirements are directly related to the objectives of the Epiwork project and are listed below.

6.2.1 Support the sharing and management of epidemiological data sets

Registered users should be able to upload annotated data sets, and a data set rating assessment mechanism should be available. The annotated data set will then compose a catalogue that will be available to users.

6.2.2 Support the seamless integration of multiple heterogeneous data sources.

Users should be able to have a unified view of related data sources. Data should be available from streaming, static and dynamic sources. All data retrieved by users or other services should be available through a common interface.

6.2.3 Support the creation of a virtual community for epidemic research.

The platform will serve as a forum for discussion that will guide the community into uncovering the necessities of sharing data between providers and modellers. Users will become active participants, generating information and providing data for sharing and collaborating online.

6.2.4 Distributed Architecture.

The Epidemic Marketplace should implement a geographically distributed architecture deployed in several sites. The distributed architecture should provide improved data access performance, improved availability and fault-tolerance.

6.2.5 Support secure access to data.

Access to data should be controlled. The marketplace should provide single sign on, distributed federated authorization and multiple access policies, customizable by users. All Epidemic Marketplace sites should rely on a common, distributed authentication infrastructure.

6.2.6 Support data analysis and simulation in grid environments.

The Epidemic Marketplace will provide data analysis and simulation services in a grid environment. Therefore, the Epidemic Marketplace should operate seamlessly with grid-specific services, such as grid security services, information services and resource allocation services.

6.2.7 Workflow.

The platform should provide workflow support for data processing and external service interaction. This requirement is particularly important for those services that retrieve data from the Epidemic Marketplace, process it, and store the processed data back in the marketplace, such as grid-enabled data analysis and simulation services.

6.3 Hardware requirements

The several EM modules will be available online and should be prepared to be used by a large user community and to manipulate large datasets. An important requirement is that this system is stable and is available at all time, so it should have a powerful and stable hardware base.

For the local deployment we have analyzed the requirements and decided for the use of two servers, complemented with two network storage units. This system provides a good level of redundancy and makes crash recovery possible in a quick and easy way. Also at first this setting allows the use of one server for testing purposes while keeping the other with a stable running version.

Connectivity should be provided initially by a shared multi-gigabit link between University of Lisbon and GÉANT, the European network for research and high

education. The access to GÉANT will assure connectivity between the main site of the Epidemic Marketplace and other Epiwork partners.

6.4 Non-functional Requirements

The main non-functional requirements that have been identified for the Epidemic Marketplace are listed below:

6.4.1 Interoperability

The Epidemic Marketplace must interoperate with modules being developed by other WPs, such as the influenza monitoring platform of WP5 and the simulation platform of WP4. In the future, systems developed by researchers across the world may need to query the EM catalogue or access the datasets available in the Epidemic Marketplace, so the seamless communication with clients based on different technologies should be supported.

6.4.2 Modularity

Not all sites where the Epidemic Marketplace will be deployed need to have all modules installed. For instance, some sites may not need a new data collector module. Therefore modularity is an important requirement of the epidemic marketplace.

6.4.3 Open-source

All software packages to be used in the implementation and deployment of the Epidemic Marketplace should be open source, as well as the new modules developed specifically to the Epidemic Marketplace. An open-source based solution reduces development cost, improves software trustworthiness and reliability and simplifies support.

6.4.4 Standards-based

In order to guarantee interoperability among the Epidemic Marketplace and software developed by other WPs, as well as the seamless integration of all geographically

dispersed sites of the Epidemic Marketplace, the system should be build over web services, authentication and metadata standards.

6.5 Repository Requirements

The Repository stores and preserves collections of data, to be interactively provided by platform users and automatically retrieved by the Collector module.

6.5.1 Separation of data and metadata

An important architectural feature for scientific repositories in general, and also the Epidemic Marketplace, is a clear separation between data and metadata (Stolte et al. 2003). For instance, there should be a clear separation between metadata and actual data schemes, since metadata may contain information not directly available in data schemes.

6.5.2 Support for Meta-data standards

Extensive support for metadata standards for web resources management and processing (e.g. searching) is required. This means the adoption of Dublin Core (Dublin Core 2009).

It is possible that only the metadata of some data sources is available through the Epidemic Marketplace, due to privacy constraints. In those cases the client should retrieve the data directly from the site hosting the data source, following directives described in the Epidemic Marketplace.

6.5.3 Ontology support

One step further in the deployment of the Epidemic Marketplace is to have a semantically enabled repository using ontologies for describing and structuring the data (Goni et al. 1997, Fox et al. 2006).

The Epidemic Marketplace will provide a framework for the creation and development of epidemiological ontologies, openly addressing the needs of this

community and fostering its active involvement. We have started using existing ontologies, such as the Unified Medical Language System (UMLS) (Bodenreider 2004).

A geographic ontology, with World coverage, is also a primary need (Chaves et al. 2005). Our goal is to contribute to making ontologies widely accepted by the Epidemiological community and ensuring their sustainable evolution, by replicating the success of similar initiatives, such as the Gene Ontology in Molecular Biology (Ashburner et al. 2000).

6.6 Mediator Requirements

The Mediator is responsible for communicating with:

1) clients, which retrieve the data collections of the Epidemic Marketplace and produce dynamical trends graphs or geographical maps according to user interaction;

2) Epiwork applications, such as Internet-based Monitoring Systems (IMS) or computational platforms (CP) for simulating the propagation of diseases;

3) other data providers, such as online news wires, RSS feeds, ProMED Mail, validated official alerts (WHO) and other event generators.

The main requirements of the mediator are:

6.6.1 Heterogeneous datasets query and search capabilities

The Mediator should manage the access to heterogeneous data from different sources, for different diseases, and in different formats, thorough either query or search interfaces.

Besides medical information, other types of information, such as geographic, sociological and related to transportation networks, need to be accessed to simulate epidemics. The heterogeneity of data depends on several factors, such as their types, representation formats, and the disease under consideration.

The data needed for an epidemic study can change significantly from disease to disease and even between studies on the same subject, depending on experimental

conditions and data collection methods. Therefore, it is important to define common data and metadata schemes enabling effective data access and analysis in the information integration processes. To achieve this goal, the data structures of a wide range of sources (e.g. IMS and CP) will be characterised for the creation of a canonical models for commonly used data sets.

6.6.2 RESTful interface

Clients should be able to search and query datasets and corresponding metadata through a RESTful interface. A RESTful interface provides ease of use, flexibility and simplicity.

6.6.3 Distributed authentication support

In order to access the Mediator, clients first should authenticate to at least one site of the Epidemic Marketplace. Authentication credentials should be shared among Epidemic Marketplace sites, and any site of the Epidemic Marketplace should access the same set of credentials for a given client.

6.6.4 Access to "plug-in-able" resources

One important feature to be supported by the Epidemic Marketplace, in particular for external data resources, is access to "plug-in-able" resources (Kuliberda et al. 2006). These external resources provide data not stored in an internal repository and may require virtualized access. Some resources can appear and disappear unexpectedly, due, for instance, to web site unavailability. "Plug-in-able" resources enable the dynamic addition of data sources through wrappers that assure physical connection to a source and convert the gathered data to one or more of the canonical data models supported by the repository.

6.7 Collector Requirements

Recent epidemiological surveillance projects are collecting data from the Internet to identify disease propagation. These systems mainly collect data from pre-selected data sources somehow related to the subject. However, other sources, like social networks

and search engine query data, may present early evidence of an infection event and propagation (Ginsberg et al. 2008). Given the increasing popularity of social networks we can find a large amount of personal information in real time, which can guide us to early detect the beginning or the propagation of an epidemic event. The main requirements of this module are:

6.7.1 Active data harvesting

To support the seamless integration of multiple heterogeneous data sources available on the Internet, the Collector should harvest information about putative infections *actively* by automatically retrieving infection alerts from the Internet through web crawlers.

6.7.2 Passive data collection

The data collector should also be able to retrieve information *passively,* by receiving information directly from online users or data providers. This can be done, for instance, through newsfeed or e-mail subscription services.

6.7.3 Local storage capability

All data collected by the data collector should be physically stored in at least one site of the Epidemic Marketplace. This is important since collected data may not be available from its original source after a predefined amount of time. Data should be stored either as datasets or in dedicated databases. Collected data should be available to clients though the mediator.

6.8 Forum Requirements

The Epidemic Marketplace will serve as an exchange platform for connecting modellers who search for data for deriving their models and those who have data who are searching for the help of modellers on interpreting their data. Therefore, another main component of the Epidemic Marketplace is the forum for discussions about the data collections and to uncover the data sharing requirements among providers and modellers.

This will help collaborations to evolve, through direct trustful sharing of data within the communities. We will guide these discussions into the direction of providing consensus agreements between modellers and data providers. The results will be reported to EU-agencies, such as the ECDC and the EMCDDA, as a contribution to setting European standards for sharing epidemic data. The main requirements of Epidemic Marketplace's forum are:

6.8.1 Group-oriented discussions

Every discussion should be associated with a group of users. Discussion participation is only granted to members of the respective group. A group-access list should be created by the user who uploads the dataset or initiates the discussion.

6.8.2 Support to distributed authentication

As it is the case with the Mediator, clients must authenticate to at least one site of the epidemic marketplace to access the forum. Authentication credentials should be shared among Epidemic Marketplace sites, and the same set of credentials for a given client should be accepted by any instance of the Epidemic Marketplace. The client will then be redirected to the Epidemic Marketplace's site which hosts the corresponding discussion, if the user is included in the corresponding group access list.

7 Conclusions and Future Work

The method of reverse engineering published epidemic modelling studies has shown to be very useful for the modelling of the metadata catalogue, since it makes it possible to understand from the analysis of the different kinds of studies, the variety of data used in epidemic studies and how it is related. This can indeed help in the development of a rich metadata model capable of describing epidemic datasets from different kinds of studies with increasing levels of accuracy.

The development and implementation of the catalogue is tightly connected with the population of the repository with different kinds of datasets and discussions within the consortium for better understanding how a metadata description can be made as exact and complete as needed and still be easily used by the occasional visitor, who deposits a dataset or wants to annotate it.

This first prototype version, based in the 15 legacy properties of DCMI, is intended as a "proof-of- concept", but it already shows the limitations spanning from annotating the datasets as described above. Some of these problems are the difficulty to describe the dataset in detail using controlled languages. With DCMI properties, the only way to annotate in detail the resources is using free text in the description field, which is a bad method, since it can be very subjective and very difficult to standardize. For this kind of informal annotation, we will provide a much simpler model, inspired on web2.0 "tags" with which the EM users can freely annotate their datasets using their own terminologies (also dubbed as "folksonomies").

We are now working on the extension of this model, using the recent DCMI terms and other specific extensions. To do so, we are working in a specific DC application model for the Epidemic Marketplace repository, where the specific extensions needed are being identified, as well as the identification of controlled languages that can be used in order to avoid subjectivity and implement a high standardization level.

7.1 Strategies for Populating the Epidemic Marketplace

We will soon have an instance of the classic "chicken-and-egg problem" in our hands, where the EM prototype is not an attractive resource because it has not a rich collection of datasets, and hasn't more datasets because the community does not perceive its potential,

The Epiwork partners who have been active in creating models using real world data, especially those directly involved in WP4, will have a key role.

Another strategy involves the active collection of data and updates to datasets from the web, for automatic annotation in the EM Catalogue or archival into the EM repository. A preliminary prototype is now automatically collecting data from Twitter on a daily basis. Figure 9 shows the number of twits (short messages) collected in one month containing the keyword H1N1 (and the alias "swine flu"), combined with one of the locations: France, Holland, Italy, Portugal and Spain. These statistics show that these publicly available sources contain relevant epidemiological data, which could not be found elsewhere on a real-time basis. It is worth noting that the Epidemic Marketplace not only collects the data, but also stores them. This is important because messages in Twitter are only available for one month. As we are periodically assembling these messages into semantically annotated data collections in the Repository, they could become a useful resource for researchers modelling the spreading of diseases. In the future we could correlate the predictions made from the data in these collections with official statistics and assess its accuracy. Previous work with web search logs data, which are private, has shown how effective these short texts can be for predicting epidemic outbreaks when the date and location of their authors can be traced.

7.2 Catalogue Implementation Calendar

The first version of the scheme is based on the legacy DC model, using the standard 15 elements. The files that are submitted to the repository are being annotated using this scheme since the repository is available online (May, 2009).

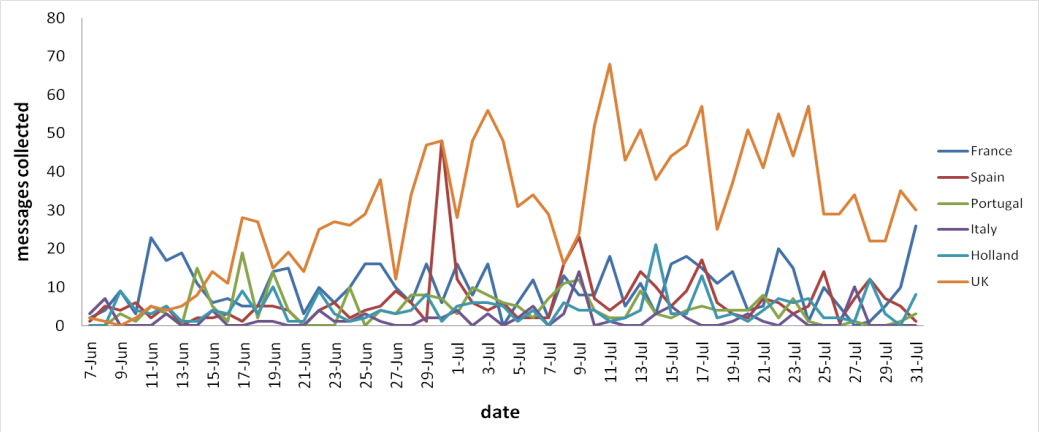

Figure 9- Number of daily collected twits with the word H1N1 in five countries.

Until January 2010 we intend to develop a metadata schema specific for the epiwork repository that is capable of supporting the correct and detailed annotation of the datasets available.

After the model is totally developed it will be necessary to implement it in the Epidemic Marketplace repository, providing new annotation tools and forms. This will be done during 2010.

8 References

Ahmadian L, De Keizer NF, Cornet R. (2009). The Use of SNOMED CT for Representing Concepts Used in Preoperative Guidelines. Stud Health Technol Inform.; 150:658-62.

Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J, Harris M, Hill D, Issel-Tarver L, Kasarskis A, Lewis S, Matese J, Richardson J, Ringwald M, Rubin, Sherlock G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25(1):25-29.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 32(Database issue):D267-270.

Bodenreider, O. & Stevens, R. (2006). Bio-ontologies: current trends and future directions. Briefings in Bioinformatics, 7(3):256-274.

Chaves M, Silva MJ, Martins B. (2005). A Geographic Knowledge Base for Semantic Web Applications. Proc. 20th Brazilian Symp. on Databases - SBBD.

Cohen, J.M, Ernst, K.C., Lindblade K.A., Vulule J.M., John C.C. and Wilson M.L. (2008). Topography-derived wetness indices are associated with household-level malaria risk in two communities in the western Kenyan highlands. Malaria J., 7: 40.

DCMI Usage Board (2008). DCMI Metadata Terms. http://dublincore.org/documents/dcmi-terms/. Accessed Sep. 25, 2009.

Dublin Core Metadata Initiative web site. http://dublincore.org/. Accessed Aug. 21, 2009.

East I.J., Hamilton S. and Garner M.G. (2008). Identifying areas of Australia at risk of H5N1 avian influenza infection from exposure to migratory birds: a spatial analysis. Geospatial health 2(2):203-213.

European Commision. http://inspire.jrc.it. Accessed Sep. 25, 2009.

Feigenbaum L., Herman I., Hongsermeier T., Neumann E., and Stephens S. The semantic web in action. Scientific American Magazine, December 2007.

Fox P, McGuinness D, Middleton D, Cinquini L, Anthony Darnell J, Garcia J, West P, Benedict J, Solomon, S. (2006). Semantically-Enabled Large-Scale Science Data Repositories. Proc. 2006 Intern. Semantic Web Conf. (ISWC), LNCS vol. 4273.

Getty Thesaurus of Geographic names online. http://www.getty.edu/research/conducting_research/vocabularies/tgn/. Accessed Sep. 25, 2009.

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. (2008). Detecting influenza epidemics using search engine query data, Nature, Letters to Editor.

Goni A, Mena E, Illarramendi A. (1997). Querying Heterogeneous and Distributed Data Repositories using Ontologies. Proc. 7th European-Japanese Conf. on Information Modelling and Knowledge Bases (IMKB'97).

Gruber, T. (1991) The role of common ontology in achieving sharable, reusable knowledge bases. Allen JF, Fikes R, Sandewall E (eds). Proceedings of KR'1991: Principles of Knowledge Representation and Reasoning. San Mateo, California: Morgan Kaufmann, pp. 601–602.

Gruber T.R. (1993). Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal Human-Computer Studies 43, p.907-928.

Harpring, P. (1997) Proper words in proper places: The thesaurus of geographic names. MDA Information, 2, pp. 5-12.

Hay, D.C. (2006). Data model patterns: A metadata map. Morgan Kaufmann, San Francisco.

IAMA- Internet Assigned Numbers Autorithy. http://www.iana.org/assignments/media-types/index.html. Accessed Sep. 25, 2009.

Kuliberda K, Blaszczyk P, Balcerzak G, Kaczmarski K, Adamus R, Subieta K. (2006). Virtual Repository Supporting Integration of Pluginable Resources. Proc. IEEE 17th Intern. Conf. on Databases and Expert Systems Applications (DEXA'06).

Luis F. Lopes, João M. Zamite, Bruno C. Tavares, Francisco M. Couto, Fabrício Silva and Mário J. Silva. (2009). Automated Social Network Epidemic Data Collector. INForum informatics symposium.

McGuinness, D.L. and Van Harmelen, F. and others (2004). Owl web ontology language overview. W3C recommendation, 10.

METeOR, Australian Institute of Health and Welfare - Metadata Online Registry. http://meteor.aihw.gov.au/content/index.phtml/itemId/181162. Accessed Sep. 25, 2009.

NCI Wiki, US National Cancer Institute - Cancer Data Standards Repository (caDSR). https://wiki.nci.nih.gov/display/caDSR/caDSR+Content. Accessed Sep. 25, 2009.

OAI-PMH (2008). The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 2.0 of 2002-06-14. http://www.openarchives.org/OAI/openarchivesprotocol.html. Accessed Sep. 25, 2009.

OAI-ORE (2008). The Open Archives Initiative Protocol for Metadata Harvesting. ORE Specification - Abstract Data Model. http://www.openarchives.org/ore/1.0/datamodel.html. Accessed Sep. 25, 2009.

Ohno-Machado L, Boxwala A, Ehresman J, Smith D, Greenes R. (1997) A Virtual Repository Approach to Clinical and Utilization Studies: Application in Mammography as Alternative to a National Database. Proc. 1997 AMIA Annual Symposium.

Powell A, Johnston P., Baker T. (2008). Domains and Ranges for DCMI Properties. http://dublincore.org/documents/2008/01/14/domain-range/. Accessed Sep. 25, 2009.

USHIK, US Health Information Knowledgebase. http://www.ushik.org/registry/index.html?Referer=Index. Accessed Sep. 25, 2009.

Starr J.M., Campbell A., Renshaw E., Poxton I.R., and Gibson G.J. (2009). Spatio-temporal stochastic modelling of *Clostridium difficile*. The Journal of Hospital Infection 71(1):49-56.

Stevens, R., Goble, C. & Bechhofer, S. (2000). Ontology-based knowledge representation for bioinformatics. Briefings in Bioinformatics, 1(4):398-414.

Stolte E, von Praun C, Alonso G, Gross T. (2003) Scientific Data Repositories – Designing for a Moving Target. Proc. ACM SIGMOD Intern. Conf. on Management of Data.

Wick, M. and Becker, T. (2007). Chapter 10: Enhancing RSS feeds with extracted geospatial information for further processing and visualization in The Geospatial Web, 2:105-116, Springer.